# Understanding Principal Components Analysis for spectral analysis



## Preamble

The goal of this at this point is to understand Principal Components Analysis (PCA) in scikit-learn.decomposition.

I have done this by writing a sequence of tests that helped me understand what was going on.

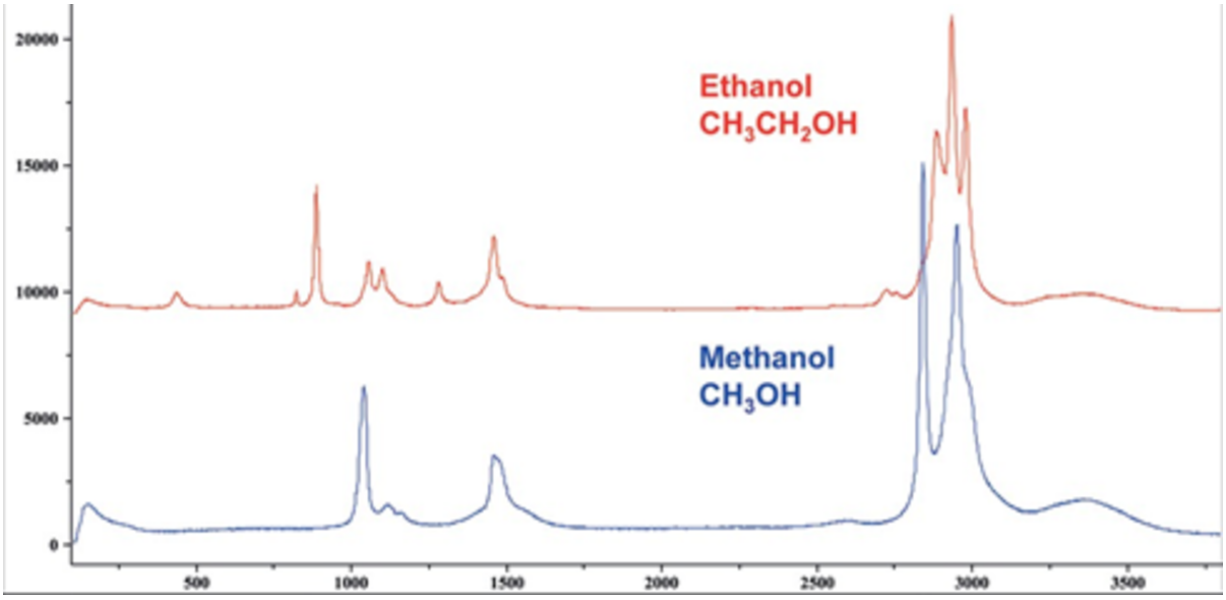
This document is a work in progress and sections will improve with my understanding.

## Note

Use Typora to read this document to see the equations.

## Introduction

Spectroscopy is the optical method of choice to detect substances or identify tissues. We have learned from very early on that identifying the peaks in a spectrum can be done to infer the substances in our sample. For instance, if we have pure substances, it is relatively easy to identify Ethanol and Methanol with their Raman spectra, because their shapes are significantly different and many peaks do not overlap:




It may even be possible to separate both substances if we have a mixture of the two, by fitting $c_e S_(\nu)_e + c_m S(\nu)_m$ to find the appropriate concentrations that can explain our final combined spectrum. However, what do we do if we have a mixture of several solutions? What if several peaks overlap? What if we don't know the original spectra?

We know intuitively that if peaks belong to the same molecule, they should vary together. If by chance none of the peaks from the different analytes overlap, then it becomes trivial: we only need to identify the peaks, find their amplitudes, and we will quickly get the concentrations of the respective analytes. But things get complicated if they have overlapping peaks, and even worse if we have more than a few components.

# Spectra as vectors

From a mathematical point of view, we can consider a spectrum as a **vector** of intensities:

$$\mathbf{I} = \sum_{i=1}^{n} I_i \hat{\nu}_i, \tag{1}$$

where each individual frequency $\nu_i$ is in its own dimension, with $\hat{\nu}_i$ the base vectors and $I_i$ is the intensity at that frequency. Therefore, if we have N=1024 points in our intensity spectrum $\mathbf{I}$, we are in an N-dimensional space, with the components being $(I_1, I_2, \dots I_N)$, and we should assume (at least for now) that these components are all independent. If we define the **norm** of a vector from the **dot product**, we can say that the norm is equal to:

$$|\mathbf{I}|^2 = \sum_{i=1}^{n} I_i \hat{\nu}_i \cdot \sum_{j=1}^{n} I_j \hat{\nu}_j = \sum_{i=1}^{n} \sum_{j=1}^{n} I_i I_j \, \hat{\nu}_i \cdot \hat{\nu}_j = \sum_{i=1}^{n} |I_i|^2, \tag{2}$$

since the spectral base vectors $\hat{\nu}_i$ are all orthonormal, which we can use to normalize a spectrum (or a vector). Finally, it is very convenient to work with matrix notation to express many of these things. We can express the spectrum $\mathbf{I}$ in the basis $\{\hat{\nu}_i\}$ with:

$$\mathbf{I} = \sum_{i=1}^{n} I_i \hat{\nu}_i = (\hat{\nu}_1\ \hat{\nu}_2\ \dots \hat{\nu}_n)\,(I_1\ I_2\ \dots\ I_n)^T = (\hat{\nu}_1\ \hat{\nu}_2\ \dots \hat{\nu}_n) \begin{pmatrix} I_1 \\ I_2 \\ \dots \\ I_n \end{pmatrix} \tag{3}$$

If we consider these matrices as partitions, we can write in a form even more compact as:

$$\mathbf{I} = \hat{\nu}[I]_\nu \tag{4}$$

where the notation $[I]_\nu$ means "the intensity coefficients in base $\nu$ to multiply the base vectors $\hat{\nu}$ and obtain the vector (spectrum)". We will use the transpose notation to keep expressions on a single line when needed.

Note that the vector itself $\mathbf{I}$ is different from the *components of that vector in a given basis* $[I]_\nu$. For more information about the notation for vector, base vectors and coefficients:

- Read Greenberg Section 10.7 on bases and base changes.
- Watch the video (in French) that explains in even more details where this comes from.
- Watch an example (in French) for problem 10.7.1 that discussed an application of this notation and formalism to perform a base change.

# Spectra as *dependent* vectors

However, we know from experience that in a spectrum, intensities are not completely independent: for instance, in the methanol spectrum above, the peak around 1000 cm$^{-1}$ has a certain width and therefore those intensities are related and are not independent. In fact, for the spectrum of a single substance, *all intensities* are related because they will come from a scaled version of the original spectrum. Therefore, if we have the reference methanol spectrum for a unity concentration $\hat{\mathbf{s}}_M$:

$$\hat{\mathbf{s}}_M = \sum_{i=0}^{n} I_{M,i} \hat{\nu}_i, \tag{5}$$

where $I_{M,i}$ is the relative intensity at frequency $\nu_i$. Any other solution of methanol of scalar concentration $c_M$ would simply yield the spectrum:

$$\mathbf{I} = c_M \hat{\mathbf{s}}_M = c_M \sum_{i=0}^{n} I_{M,i} \hat{\nu}_i. \tag{6}$$

So if we have several individual solutions $\{\hat{\mathbf{s}}_j\}$ from which we create a mixture of concentrations $c_j$, we will generate spectra in a sub-space of the original $n$-dimensional intensity vector-space. The set of vectors $\{\hat{\mathbf{s}}_j\}$ is a basis set because we can generate all vectors in that sub-space with a linear combination of the vectors (or spectra). The dimension of that subspace is equal to the number of elements in $\{\hat{\mathbf{s}}_j\}$ We can write the mixture spectrum $\mathbf{I}$ as:

$$\mathbf{I} = \sum_j c_j \hat{\mathbf{s}}_j = (\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \dots, \hat{\mathbf{s}}_n)\,(c_1, c_2, \dots, c_n)^T = \hat{\mathbf{s}}[c]_{\hat{\mathbf{s}}} \tag{7}$$

Again, we read the last expression $[c]_{\hat{\mathbf{s}}}$ as "the coefficients in base $\{\hat{\mathbf{s}}\}$ needed to multiply the base vectors $\hat{\mathbf{s}}_i$ to obtain the final spectrum $\mathbf{I}$". It stresses the point that the vector $\mathbf{I}$ and its components in a given basis $[c]_{\hat{\mathbf{s}}}$ are not the same thing. This will become critical below when we look a Principal Components.

Finally, if we want to describe a **collection** of $m$ spectra obtained from mixing these base solutions $\hat{\mathbf{s}}$ with concentrations $c_{ij}$ for the i-th spectrum and the j-th base solution, we can write:

$$\mathbf{I}_i = \sum_j c_{ij} \hat{\mathbf{s}}_j. \tag{8}$$

This can be rewritten in matrix notation:

$$(\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_m) = (\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \ldots, \hat{\mathbf{s}}_n) \begin{pmatrix} c_{11} & c_{21} & \ldots & c_{m1} \\ c_{12} & c_{22} & \ldots & c_{m2} \\ \ldots & \ldots & \ldots & \ldots \\ c_{1n} & c_{2n} & \ldots & c_{mn} \end{pmatrix} \tag{9}$$

to yield this very compact form:

$$\mathbf{I} = \hat{\mathbf{s}}[\mathbf{C}]_{\hat{\mathbf{s}}} \tag{10}$$

This equation represents, in a single expression, the $m$ spectra obtained by mixing the $n$ solutions with concentrations $c_{ij}$ for the $i$-th spectrum and the $j$-th solution.

# Final notes on intensity spectra as vectors

If we have several components (i.e. methanol, ethanol, etc...) and there is no overlap whatsoever between the spectra (i.e the peaks are all distinct), then the base vectors $\hat{b}_i$ and $\hat{b}_j$ are orthogonal. However, it is more likely that the solutions *do* have overlapping spectra, therefore the base vectors (and consequently the base itself) will *not be orthogonal*. It is perfectly acceptable to have a base that is not orthogonal: it remains a base because any linear combination can create any spectrum we would measure.

# Base change

The general expression for a vector as a function of its basis and its components in that basis is such that obviously, it stands correct for any basis:

$$\mathbf{I} = \hat{\mathbf{b}}[\mathbf{C}]_{\mathbf{b}} = \hat{\mathbf{b}}'[\mathbf{C}']_{\mathbf{b}'} \tag{11}$$

It is the purpose of the present section to show how to go from a basis b to a basis b', that is, how to transform the coefficients c into coefficients c'. For more information, you can look at the Youtube Video on base changes.

Since we can express any vector in a basis, we can choose to express the basis vectors $\hat{\mathbf{b}}$ in the $\hat{\mathbf{b}}'$ basis, with the coefficients $[\mathbf{Q}]_{\hat{\mathbf{b}}'}$ we do not know yet:

$$\hat{\mathbf{b}} = \hat{\mathbf{b}}'[\mathbf{Q}]_{\hat{\mathbf{b}}'}, \tag{12}$$

where each column of the matrix $[\mathbf{Q}]_{\hat{\mathbf{b}}'}$ is the component of the vector $\hat{b}_i$ in the $\hat{b}'$ basis. By definition, a basis set has enough vectors to cover the vector space, therefore both basis sets must have the same number of vectors, and the matrix $[\mathbf{Q}]_{\hat{\mathbf{b}}'}$ is necessarily square, and can be inverted. We can therefore use (11) in (12) and obtain simply:

$$\mathbf{I} = \hat{\mathbf{b}}[\mathbf{C}]_{\mathbf{b}} = \left(\hat{\mathbf{b}}'[\mathbf{Q}]_{\hat{\mathbf{b}}'}\right)[\mathbf{C}]_{\mathbf{b}} = \hat{\mathbf{b}}'\left([\mathbf{Q}]_{\hat{\mathbf{b}}'}[\mathbf{C}]_{\mathbf{b}}\right) = \hat{\mathbf{b}}'[\mathbf{C}']_{\mathbf{b}'} \tag{13}$$

This means that, when the vectors in different bases are expressed by (11), the coordinates in the basis $\hat{b}'$ can be obtained from the components in the basis $\hat{b}$ by this simple transformation:

$$[\mathbf{C}']_{\mathbf{b}'} \equiv [\mathbf{Q}]_{\hat{\mathbf{b}}'}[\mathbf{C}]_{\mathbf{b}} \tag{14}$$

## PCA base change in `sklearn`

Equation (11) is not the only possibility to express a vector in different basis. We will see later that Principal Component Analysis (PCA) often *translates* the sample vectors (i.e. the intensity spectra) to the "origin" by subtracting the mean spectrum from all spectra. This means that we have a more general transformation than (11) in that we do not express $\mathbf{I}$ in a different set of coordinates but rather $\mathbf{I} - \bar{\mathbf{I}}$:

$$\mathbf{I} - \bar{\mathbf{I}} = \hat{\mathbf{b}}'[\mathbf{C}']_{\mathbf{b}'}, \tag{15}$$

with

$$\bar{\mathbf{I}} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{I}_i \equiv (\hat{\nu}_1, \hat{\nu}_2, \ldots, \hat{\nu}_n)\left(\bar{\mathbf{I}}_1, \bar{\mathbf{I}}_2, \ldots, \bar{\mathbf{I}}_n\right)^T \equiv \hat{\nu}\left[\bar{\mathbf{I}}\right]_{\nu} \tag{16}$$

This average is computed in the "intensity" basis (or original basis $\{\hat{\nu}\}$) because that is the only basis we know when we start (i.e. we average all spectra). This small change where we subtract the mean is important, because to return to another basis used to generate $\mathbf{I}$, we need to write:

$$\mathbf{I} = \hat{\mathbf{b}}'[\mathbf{C}']_{\mathbf{b}'} + \bar{\mathbf{I}} = \hat{\mathbf{b}}[\mathbf{C}]_{\mathbf{b}}, \tag{17}$$

and if we try to follow the same development as in (13), we would quickly get stuck because we do not have $\bar{\mathbf{I}}$ neither in $\{\mathbf{b}\}$ or $\{\mathbf{b}'\}$ coordinates, we have it in the $\{\hat{\nu}\}$ coordinates :

$$\mathbf{I} = \hat{\mathbf{b}}'[\mathbf{C}']_{\mathbf{b}'} + \hat{\nu}\left[\bar{\mathbf{I}}\right]_{\nu} = \hat{\mathbf{b}}[\mathbf{C}]_{\mathbf{b}}, \tag{18}$$

Yet, this is the situation we will encounter later:

1. $\mathbf{I}$ is the many spectra we have acquired in the lab. They are in $\{\hat{\nu}\}$ basis (i.e. simple intensity spectra).
2. We can compute $\bar{\mathbf{I}}$ with (16), also in the spectral component basis $\{\hat{\nu}\}$.
3. The $\{\mathbf{b}'\}$ basis is the Principal Component Analysis (PCA) basis that will be obtained from the module together with the coefficients $[\mathbf{C}']_{\mathbf{b}'}$. It comes from a singular value decomposition, and at this point, we do not worry oursleves with how it is obtained: we know we can obtain $\{\mathbf{b}'\}$ and $[\mathbf{C}']_{\mathbf{b}'}$ from `sklearn` and PCA.
4. Finally, the $\{\mathbf{b}\}$ basis is the "solution" basis for which we would like to get the concentrations $[\mathbf{C}]_{\mathbf{b}}$ for

our lab measurements. We know *some* $\{\mathbf{b}_i\}$, but we may not know them all. In Raman, this could be the lipid spectrum, DNA spectrum, protein spectrum etc... We want the coefficients to try to determine the concentrations of these molecules and get insight (or answers) from our experimental spectra, but we may not have all the components (i.e. we may not have the full basis).

There is mathematically not much we can do with these three coordinate systems in $(18)$, unless we express the average spectrum $\bar{\mathbf{I}}$ in one or the other bases. We can do two things:

1. Express $\bar{\mathbf{I}}$ in the base $\{\mathbf{b}\}$
2. Express $\bar{\mathbf{I}}$ in the base $\{\mathbf{b}\}'$

For reasons that should become clear later, the `sklearn` PCA python module that we will use performs the multiplication $\hat{\mathbf{b}}'[\mathbf{C}']_{\mathbf{b}'}$ to express the result in $\{\hat{\nu}\}$ space *before* adding the mean, therefore we will choose to express $\bar{\mathbf{I}}$ in the base $\{\mathbf{b}\}$ :

$$\hat{\mathbf{b}}'[\mathbf{C}']_{\mathbf{b}'} + \hat{\mathbf{b}}\left[\bar{\mathbf{I}}\right]_{\mathbf{b}} = \hat{\mathbf{b}}[\mathbf{C}]_{\mathbf{b}}, \tag{19}$$

$$\hat{\mathbf{b}}'[\mathbf{C}']_{\mathbf{b}'} = \hat{\mathbf{b}}[\mathbf{C}]_{\mathbf{b}} - \hat{\mathbf{b}}\left[\bar{\mathbf{I}}\right]_{\mathbf{b}}, \tag{20}$$

$$\hat{\mathbf{b}}'[\mathbf{C}']_{\mathbf{b}'} = \hat{\mathbf{b}}\left([\mathbf{C}]_{\mathbf{b}} - \left[\bar{\mathbf{I}}\right]_{\mathbf{b}}\right), \tag{21}$$